IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 May 2022, accepted 13 June 2022, date of publication 27 June 2022, date of current version 5 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3186444

Multiclass Classification Performance Curve

JESÚS S. AGUILAR-RUIZ[1] **AND MARCIN MICHALAK**[2]
[1]School of Engineering, Pablo de Olavide University, 41013 Seville, Spain
[2]Department of Computer Networks and Systems, Silesian University of Technology, 44-100 Gliwice, Poland

Corresponding author: Jesús S. Aguilar-Ruiz (aguilar@upo.es)

This work was supported in part by MCIN/AEI/10.13039/501100011033 under Grant PID2020-117759GB-I00; and in part by the Andalusian Plan for Research, Development and Innovation and the Department of Computer Networks and Systems (RAu9), Silesian University of Technology.

ABSTRACT Quality of predictive models is a critical factor. Many evaluation measures have been proposed for binary and multi–class datasets. However, less attention has been paid to graphical representation of the classification performance, where the ROC curve is extensively used for binary datasets but there is no standard method accepted by the scientific community for multi–class datasets. In this work, a multi–class classification performance (MCP) curve based on the Hellinger distance between true and prediction probabilities of the classifier is introduced. The MCP curve shows the classification performance, contributes to highlight the low or high confidence on correct predictions, and quantifies the quality by means of the area under the curve.

INDEX TERMS Classification, machine learning, multi–class data, performance curve, predictive models, ROC curve.

I. INTRODUCTION

Quality of decision is an important concept in machine learning, because it assesses the performance of a predictive model in terms of comparison with reality. The simplest measure of decision quality is the accuracy, i.e. the fraction of correct decisions. While it seems obvious that high accuracy is good, the truth is that it might be very misleading. Many medical datasets report very few cases of positive cases (disease) against negative cases (normal)—very common for rare diseases. For example, if only 3% of patients in the dataset have colorectal cancer, a predictive model that blindly provides negative case would be correct 97% of the time.

Accuracy is, in general, useless in various contexts, particularly in biomedicine and psychology. Therefore, when prevalence is not around 0.5 (balance between positive and negative cases) the accuracy looses credibility as a measure of quality. To redeem the issue there exist other measures that can be useful to compare the performance of several predictive models, since even when the accuracy is equal for two models, their performances could be quite different – one due to a low rate of false negative cases and the other to false positive cases. More sophisticated measures, like the Matthews correlation coefficient [1], the K–category correlation coefficient, R_K (less well–known generalization of the two–class Matthews correlation coefficient) [2], Cohen's κ [3], or the F1–score [4] try to provide a better perspective.

The interpretation of false decisions (positive or negative) could lead to complex situations, e.g. expensive medical treatments for healthy patients (false positive) or absence of treatment for seriously ill patients (false negative). Sensitivity and specificity are two measures that appear to mitigate the effect of an incorrect interpretation of accuracy. Sensitivity is the ratio between the true positive decisions and the number of positive cases; specificity is the ratio between the true negative decisions and the number of negative cases. Both are perfectly combined in a very useful graphical measure: the Receiver –or Relative– Operation Characteristic (ROC) curve.

The ROC curve [5]–[7] represents the relation between the false positive rate (FPR) and the true positive rate (TPR), and illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied from the confusion matrix. FPR and TPR are calculated from sensitivity and specificity ($FPR = 1 - specificity$ and $TPR = sensitivity$). The curve shows a solid idea of the behavior of a classifier, and allows to

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

compare several classifiers in one image for the same dataset. Also, it has been proven that the area under the ROC curve is an excellent measure of classification performance [8], as it conveys more information than many scoring metrics by visualizing the performance of the classifier by a curve instead of providing a single scalar value. Most importantly, the ROC curve, unlike the accuracy, only depends on the FPR and TPR, which are independent of imbalance, i.e. of class distribution. Given its easy interpretation and usefulness for comparative analysis, the ROC curve has been used in myriads of diverse applications [9]–[12], sometimes with questionable statistical confidence [13]. However, its binary character has substantially limited its use to strictly dichotomous decision contexts.

The ROC curve presents many interesting properties, but also an important shortcoming: it can only be applied to two–class (binary) datasets. In case of multi–class datasets (more than two labels for the nominal target variable), there is no solution that aggregates and summarizes the classifier behavior for all the classes together (e.g. 3–classes: early and advanced disease, and normal cases). The most common approach is to represent the class of interest (positive) against the others (negative), but it does not contribute to understand the global performance of the classifier (also, information is always lost in the reduction of a problem to a dichotomy). If all the classes are important, the smoothed solution named *macro–average* is to average all the ROC curves (one–versus–rest for each class), but this compromises its natural insensitivity to class skew. On the contrary, the *micro–average* approach takes into account the proportion of every class (aggregating the contributions of all classes), given more importance to larger classes and, therefore, assigning greater values when the largest class performs better (i.e. it is dominated by the more frequent class). In short, *macro–average* gives equal importance to each class, and *micro–average* to each sample. Obviously, in case of equal number of samples for each class, then both *macro–* and *micro–average* ROC curves will result in exactly the same score. However, when the imbalance is significant, both approaches lead to very different curves with refutable interpretations.

The area under the ROC curve (AU(ROC)) is a robust evaluation measure [14]. It is equivalent to the probability that a randomly chosen positive sample will be rated higher than a negative sample, and it is useful to compare classifiers when no one dominates the others [15]. There exist approaches for the AU(ROC) in the context of multi–class classification, based on averaging pairwise comparison of classes [16] or on the use of K–dimensional space to compute the volume of the ROC surface [17]–[19], being K the number of class labels. However, none of these approaches provide a graphical representation.

This paper presents an intuitive method to calculate the classification performance in the multi–class context, i.e. for datasets whose target variable contains any number of class labels. The approach provides a two–dimensional classification performance curve (MCP curve) to visualize the behavior of predictive models and subsequently, as a scalar measure of quality, the area under the MCP curve (AU(MCP)). Unlike most approaches for binary datasets, it is not based on the confusion matrix but on the prediction probabilities generated by the classifier for the K class labels.

The paper is organized as follows: after introducing the research context, Sec. II presents the mathematical foundations based on probability distributions that lead to the definition of the MCP curve; Sec. III shows, based on the results of two classifiers for multi–class data, how the MCP curve can graphically compare classifier performances, what is not possible with ROC curves; finally, Sec. IV discusses the most important conclusions and future work.

II. MULTI–CLASS CLASSIFICATION PERFORMANCE

Let $D = \{e_i | e = (\overline{x}, y), \overline{x} \in \mathbb{X}^m, y \in \mathbb{Y}, i = 1, 2, \ldots, n\}$ be a dataset with n samples, which belong to a m–dimensional feature space and have a corresponding outcome y in the space $\mathbb{Y}$. When $|\mathbb{Y}| = 2$ we say that it is a 2–class (binary) classification problem; otherwise, when $|\mathbb{Y}| > 2$, we say that it is a multi–class classification problem.

For simplicity, let us assume that $\mathbb{Y} = \{1, 2, \ldots, K\}$ (set of class labels). The true probability of sample e_i, denoted by $t(e_i)$, can be encoded as a K–dimensional one–hot vector, which all the values are 0 except for one 1 at the position k, that satisfies $y_i = k$ (following the *indicator function*, for all k, $t(e_i)_k = 1_{y_i=k}$), being $k \in \mathbb{Y}$. Obviously, there will only be K different true probability vectors. The prediction probability of sample e_i, denoted by $p(e_i)$, is also a K–dimensional vector, whose values are provided by the classifier as output probabilities of belonging to class $k \in \mathbb{Y}$.

In order to observe the quality of the classifier prediction, a distance function between the two discrete probability distributions must be applied to each sample. Let d be a distance function between two distributions t and p, such as $d : [0, 1]^K \times [0, 1]^K \rightarrow [0, 1]$ (certainly, the true probability could be expressed in the space $\{0, 1\}^K$, but it is generalized to $[0, 1]^K$ with the intention of including uncertainty in later work). The interest lies in calculating the distance d for each sample $e \in D$, i.e. $d(t(e_i), p(e_i))$ for all $i = 1, 2, \ldots, n$. The distance value is very important because it informs about how far the prediction is from the true observation for each sample. Values close to 0 mean that the prediction is good, and values close to 1 mean that the prediction substantially differs from the observed class. The value $\varphi_i = 1 - d(t(e_i), p(e_i))$ can be interpreted as the probability that the classifier correctly assigns the observed class label to the test sample e_i (*certainty*).

The Hellinger distance [20], [21], is a metric oriented to probability distributions. Let $P = [p_1, \ldots, p_K]$ and $Q = [q_1, \ldots, q_K]$ be two discrete probability distributions. The Hellinger distance is defined as:

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{K} \left(\sqrt{p_j} - \sqrt{q_j}\right)^2} \tag{1}$$

The choice of the Hellinger distance is due to its interesting properties: a) it is symmetric; b) bounded (defined in [0,1] thanks to the factor $\frac{1}{\sqrt{2}}$); c) convex with respect to both P and Q; and d) it satisfies the triangle inequality. As a consequence, it is recently attracting considerable amount of attention in many scientific fields [22]–[30].

The discrete α–divergence [31] defines a family of functions as follows:

$$D_\alpha(P||Q) = \frac{\sum_{j=1}^{K}\left(\alpha p_j + (1-\alpha)q_j - p_j^\alpha q_j^{1-\alpha}\right)}{\alpha(1-\alpha)} \quad (2)$$

$$\alpha \in \mathbb{R} \setminus \{0, 1\}$$

The α–divergence has some important properties: a) non–negative; b) convex with respect to both P and Q; c) it is 0 if and only if $P = Q$. In addition, there is a very interesting relationship between the Hellinger distance and the Kullback–Leibler divergence [32] through the α–divergence.

The special cases of $\alpha = 0$ and $\alpha = 1$ are related to the Kullback–Leibler divergence (KL):

$$\lim_{\alpha\to 0} D_\alpha(P||Q) = KL(Q||P), \quad \lim_{\alpha\to 1} D_\alpha(P||Q) = KL(P||Q) \quad (3)$$

The case of $\alpha = \frac{1}{2}$ is related to the Hellinger distance:

$$D_{\frac{1}{2}}(P||Q) = 2\sum_{j=1}^{K}\left(\sqrt{p_j} - \sqrt{q_j}\right)^2 \quad (4)$$

Hence the Hellinger distance $H(P, Q) = \frac{1}{2}(D_{1/2})^{1/2}$ can be expressed as a α–divergence, and its square is related to the midpoint between $KL(Q||P)$ and $KL(P||Q)$ taking $D_\alpha(P||Q)$ as a reference, although unlike KL–divergence, $H(P, Q)$ is a mathematically valid metric. $H(P, Q)$ can also be expressed as a function of the symmetric Bhattacharyya coefficient (BC) [33] (also known as *fidelity*), which can be derived from the Chernoff α–coefficient [31]:

$$H(P, Q) = \sqrt{1 - BC(P, Q)} \quad (5)$$

where

$$BC(P, Q) = \sum_{j=1}^{K}\sqrt{p_j q_j}$$

If the distance between the true probability $t(e_i)$ and the prediction probability $p(e_i)$ of a sample e_i is close to 0, then the classifier is making a good prediction for e_i; otherwise, when it is close to 1, the prediction is bad. Good or bad predictions are not necessarily correct and incorrect, respectively. The distance is a quality measure of the classification performance, so it would be possible to provide at the same time a not small distance associated with a correct prediction (e.g. for a 3–class problem, when $t = [1, 0, 0]$ and $p = [0.4, 0.3, 0.3]$, $H(t, p) = 0.606$ and the prediction is correct).

The boundary for a correct prediction for a K–class classification problem is given by the expression:

$$H < \theta = (1 - K^{-1/2})^{1/2} \quad (6)$$

that means all the correct predictions must satisfy the expression (it is necessary but not sufficient condition).

Let $\varphi_i = 1 - H(t(e_i), p(e_i))$ be the performance of the classification model with sample e_i. The measure φ_i is equivalent to the probability of correctly classifying the sample e_i. Therefore, if $\varphi_i \leq 1 - \theta$ then sample e_i will not be correctly classified. However, the opposite is not true, that is, even when $\varphi_i > 1 - \theta$ there will be some samples that will not be correctly classified (e.g. when $t = [1, 0, 0]$ and $p = [0.4, 0.5, 0.1]$ the prediction is not correct), thus establishing a theoretical bound for incorrect classifications (error rate). In fact, this is a very interesting aspect, because two classifiers both with accuracy $= 1$ (no errors), could vary their aggregated mean distance $\bar{H}$, so the performance would be very different in terms of prediction certainty (equal in accuracy, but not in confidence).

Mathematically, a *sequence* S is an ordered collection of objects, $(S_n)_{n\in\mathbb{N}}$. The set Φ of values φ_i calculated for all the test samples can be increasingly ordered, producing a sequence $\Phi' = (\varphi'_1, \ldots, \varphi'_n)$. The sequence Ω of points $\left(\left(\frac{i-1}{n-1}, \varphi'_i\right)\right)_{i=1}^{n}$ can be plotted as a line connecting the points, and it will provide a curve within the unit square. The AU(MCP) can also be obtained as a quantification of the classifier performance, comparable to AU(ROC) (trapezoidal areas between every two consecutive points), although in this case for multi–class datasets.

The MCP curve algorithm design (see Alg. 1) prioritizes interpretability over efficiency. In case of k–fold cross–validation, it would be necessary to join all the k sequences Φ_k and then (merge)sort the final sequence, before generating the points of Ω. The complexity of the algorithm is $O(n \log n)$ and it is independent on the complexity of the classifier.

The AU(MCP) is easily computed as follows:

$$AU(MCP) = \frac{1}{n-1}\left(\sum_{i=1}^{n}\Phi'(i) - \left(\frac{\Phi'(1) + \Phi'(n)}{2}\right)\right) \quad (7)$$

where $n = |\Phi'|$. As $\forall i\ \Phi'(i) \in [0, 1]$, since $H(t(e_i), p(e_i)) \in [0, 1]$, then $AU(MCP) \in [0, 1]$.

As Ω contains all the points of the MCP curve, increasingly ordered by φ, the first α points would not likely be greater than $1 - \theta$, which means these samples are not correctly classified. From the point α all the rest might be correctly classified, although not with the same confidence, as the values of φ would range in $[1 - \theta, 1]$. However, if $BC(t(e_i), p(e_i)) > 0.5$ then sample e_i is correctly classified, which defines another threshold for H values:

$$H > \delta = \sqrt{1 - \sqrt{0.5}} \quad (8)$$

Therefore, if $\varphi < 1 - \theta$ then the sample is incorrectly classified; if $\varphi > 1 - \delta$ then the sample is correctly classified; and if $\varphi \in [1 - \theta, 1 - \delta]$ then the behavior of the

TABLE 1. Confusion matrix, (left) and precision, sensitivity and specificity (right) from drug consumption dataset (target: heroin) for Naïve–Bayes (NB) (1a) and random forests (RF) (1b).

		Predicted									
	Class	0	1	2	3	4	5	6	Prec.	Sens.	Spec.
True	0	1086	6	7	3	19	2	482	91.67	67.66	64.64
	1	46	0	1	0	1	0	20	0.00	0.00	99.56
	2	29	2	2	0	4	0	57	15.38	2.13	99.39
	3	19	0	0	0	2	0	44	0.00	0.00	99.84
	4	3	0	2	0	2	1	16	6.67	8.33	98.50
	5	1	0	1	0	2	0	12	0.00	0.00	99.84
	6	1	0	0	0	0	0	12	1.87	92.31	66.29

(a) NB (Accuracy= $\mathbf{0.585}$; Cohen's $\boldsymbol{\kappa} = \mathbf{0.100}$).

		Predicted									
	Class	0	1	2	3	4	5	6	Prec.	Sens.	Spec.
True	0	1601	1	2	0	1	0	0	85.30	99.75	1.43
	1	66	2	0	0	0	0	0	66.67	2.94	99.94
	2	94	0	0	0	0	0	0	0.00	0.00	99.89
	3	64	0	0	1	0	0	0	50.00	1.54	99.95
	4	24	0	0	0	0	0	0	0.00	0.00	99.95
	5	15	0	0	1	0	0	0	0.00	0.00	100.00
	6	13	0	0	0	0	0	0	0.00	0.00	100.00

(b) RF (Accuracy= $\mathbf{0.851}$; Cohen's $\boldsymbol{\kappa} = \mathbf{0.019}$).

Algorithm 1 MCP Curve

Require:
D_λ, D_π: Datasets for learner (λ) and predictor (π)
C: Classifier
d: Distance function
Ensure:
Ω: Sequence of points of the MCP curve

$\Omega \leftarrow ()$ ▷ Empty sequence (MCP curve)
$\Phi \leftarrow \{\}$ ▷ Empty set (Certainty)
$M \leftarrow C(D_\lambda)$ ▷ M: Model
$p \leftarrow M(D_\pi)$ ▷ p: Prediction probability matrix
for $e_i \in D_\pi$ **do**
  $\varphi \leftarrow 1 - d(t(e_i), p(e_i))$ ▷ t: True probability vector
  $\Phi \leftarrow \Phi \oplus \varphi$ ▷ $\oplus$: Insertion at the end
end for
$\Phi' \leftarrow \text{Sort}(\Phi)$ ▷ Increasing order
for i in $\{1, \ldots, |\Phi'|\}$ **do**
  $\Omega \leftarrow \Omega \oplus \left(\frac{i-1}{|\Phi'|-1}, \Phi'(i)\right)$ ▷ $\Phi'(i)$: φ at position i in Φ'
end for

classifier is uncertain. As the MCP curve is a monotonically increasing function, the prediction performance will depend on the shapes of the function within the three possible regions (incorrect, uncertain and correct) defined by the two points $(\alpha, 1-\theta)$ and $(\beta, 1-\delta)$. Thus, given these two points, we could observe many different performance behaviors for the same confusion matrix, in which is based most of the classification performance measures. The promising scenario opened by the MCP curve requires further analysis and study of the properties in relation to the most commonly used metrics in the scientific literature, since prediction probabilities have much more potential than the confusion matrix (counters from the one–hot vector provided by a discrete mapping function over the probabilities: $g : [0, 1]^K \rightarrow \{0, 1\}^K$).

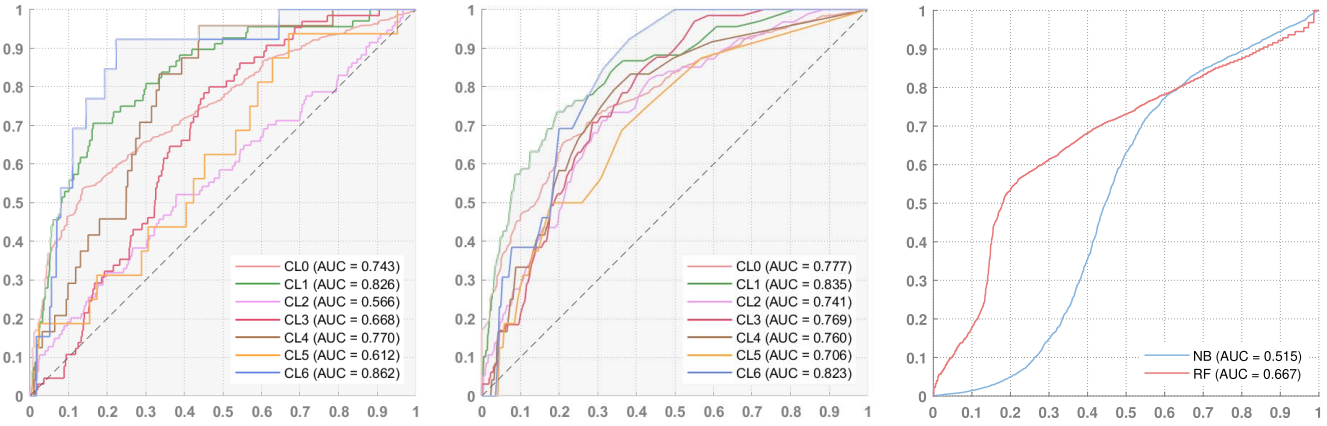
III. EXPERIMENTS

The Drug Consumption dataset [34] contains records for 1,885 respondents, and 12 attributes: neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, impulsivity, sensation seeking, level of education, age, gender, country of residence and ethnicity. All input attributes were originally categorical and were quantified by the authors to be considered as real–valued, by using ordinal and nominal feature quantification techniques (polychoric correlation and non–linear categorical principal component analysis, respectively). In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug –Semeron– which was introduced to identify over–claimers). For each drug they have to select one of the answers: never used the drug (CL0), used it over a decade ago (CL1), or in the last decade (CL2), year (CL3), month (CL4), week (CL5), or day (CL6). Therefore, the dataset contains eighteen 7–class classification problems.

Two well–known classifiers were used: Naïve–Bayes (NB) and Random Forests (RF). All the experiments were validated by means of stratified 10–fold cross–validation.

Basic statistics for NB (default parameters) and RF (gain ratio as splitting criterion for decision trees, no pruning and 500 trees) are illustrated in Tables 1a and 1b, respectively (results for accuracy and the Cohen's κ are also included). The difficulty of classification is remarkable, mainly due to the imbalance of the target variable. Values for precision and sensitivity are notably low for all the classes except for CL0, which represents about 85% of data([CL0, 85.1%], [CL1, 3.6%], [CL2, 5%], [CL3, 3.5%], [CL4, 1.3%], [CL5, 0.8%], [CL6, 0.7%]). However, the interest lies in classes with high number (CL4–CL6), as they indicate recent drug consumption. Accuracy, as overall measure, reveals a much better performance of RF (0.851) over NB (0.585). However, RF is assigning most test cases to the majority class (CL0), which is completely useless. On the other hand, the Cohen's κ (measure of agreement between what it is relatively observed and what it would be expected by chance) does show the unsatisfactory behavior of NB (0.100) and, more importantly, of RF (0.019), as opposed to accuracy [35].

Considering the ROC curves, which in principle should visually show the behavior of each class with respect to the others, the interpretation becomes inconsistent and inconclusive. For NB (Fig. 1a), the behavior ranges from the worst (CL2) to the best (CL6) over a wide range. However, the accuracy for CL6 is minimal (1.87% in Tab. 1a). For RF (Fig. 1b), the behavior is more stable, with lower variance, but



(a) One–versus–all ROC curves for NB. (b) One–versus–all ROC curves for RF. (c) The MCP curve for NB and RF.

FIGURE 1. Graphs for drug consumption dataset (target: heroin) by using two classifiers: Naïve–Bayes (1a) and random forests (1b). The MCP curve (1c) jointly shows the performance of both classifiers ($1-\theta=0.21$ and $1-\delta=0.46$). The areas under the curves (AUC) are also shown.

still offers good performance for CL6 ($AU(ROC)=0.823$), although its precision is null. In sum, both accuracy and ROC curves (as well as the area under the curve), offer quality indicators that tend to overestimate the real performance of the classifiers.

The curves depicted in Fig. 1c shows the multi–class classification performance (MCP) curve of Naïve Bayes and Random Forests classifiers on the Drug Consumption dataset for the target heroin. The best case would be when all the distances are zero, so the φ values would be 1, and the curve would be a flat line equivalent to the constant function $y=1$ (AU(MCP) would be 1). The worst case would be when all the distances are 1, i.e. the true and prediction probabilities are maximally distant, and the curve would be a flat line equivalent to the constant function $y=0$ (AU(MCP) would be 0). Therefore, the visual interpretation of the MCP curves is quite similar to the ROC curve. The goal of this image is to show that both the shape of and the AU(MCP) consistently illustrate the classification performance of both predictive models (NB and RF).

RF reaches quickly the values $1-\theta=0.21$ (for $x=0.12$) and $1-\delta=0.46$ (for $x=0.16$), which means that 12% and 84% of samples are incorrectly and correctly classified, respectively, and only 4% of uncertainty. For NB, $x=0.33$ and $x=0.43$, respectively, which means 33% and 57%, and 10% of uncertainty. These values are directly related to the accuracy ($NB=0.585$ and $RF=0.851$) shown in Tabs. 1a and 1b. As for the values of AU(MCP) ($NB=0.515$ and $RF=0.667$), they are more conservative than the values of accuracy, which is consistent with the very low precision and sensitivity values for six out of seven class labels. The most relevant aspect is that it is observed in the MCP curve the low confidence of predictions for correctly classified cases, as both classifiers only provide probabilities close to 1 for just a few sample cases. In contrast, the ROC curves (Figs. 1a and 1b) reach high TPR values very soon (when $FPR=0.5$, values of TPR are over 0.8 for RF).

Considering the accuracy and ROC results for RF, it could be concluded that they are satisfactory, although they are certainly not reliable. The MCP curve does not only show the classification performance, but also contributes to highlight the low or high confidence on correct predictions.

There exists a structural difference between ROC and MCP curves: the ROC curve always starts at $(0,0)$ and ends at $(1,1)$, because it is consequence of setting the decision threshold at 1 and 0, respectively. However, the MCP curve could start at any point $(0,\varphi_{min})$ and ends at any point $(1,\varphi_{max})$, being φ_{min} and φ_{max} the minimum and maximum values of distance between the true and the prediction probabilities of samples, respectively. For instance, if none sample e_i gets a perfect prediction, with $d_i=0$, then $\varphi_i<1$, and will not reach the upper–right corner of the unit square. This feature is also interesting when comparing the performance of classifiers.

IV. CONCLUSION

Multi–class classification is a very common and important problem. Many quality measures, most of them based on the confusion matrix, exist to assess classification performance for binary data sets. The ROC curve stands out from the others in that it provides a graphical representation of classifier performance, and also a quantification of its quality (AU(ROC)). However, it is not possible for it to provide a unique representation for multi–class datasets. Several approaches have been formulated, like micro–average and macro–average ROC curves, which cannot provide sufficient information in metric multi–class classification efficiency in any scenarios. The MCP curve arises as an alternative to fill this gap in the context of multi–class classification.

The confusion matrix is based on prediction probabilities. All the values of this matrix are calculated by the *arg max* function from the probabilities (class with the largest predicted probability). Therefore, real numbers (probabilities) are transformed into a one–hot vector for each test instance

before updating the confusion matrix. The MCP curve works directly with prediction probabilities, avoiding a loss of information (with respect to classification performance) caused by the transformation into the confusion matrix.

Understanding the confidence of predictions is an important issue, because it lends mathematical continuity to discrete decisions. To calculate the MCP curve, the classic correct or incorrect classification is replaced by probabilities, which enriches the interpretation of results and allow enhance the behavior of predictive models. For example, for a 4–class problem, a medical decision would not be as reliable when the probability for the majority class is 1 as it would be when it is 0.51 (both correct). In well–known measures like accuracy, the Matthews correlation coefficient or F1–score, the decision is made by voting without analyzing the outcome of prediction probabilities for each class. Therefore, from the perspective of reliability, for a given dataset, several classifiers yielding exactly the same confusion matrix might provide different MCP curves, which contributes to a deeper insight into the prediction performance of each classifier.

As stated, the ROC curve is very useful to graphically compare the prediction performance of classifiers when datasets are binary (2–class), but it is not applicable to multi–class datasets. The MCP curve offers to the research community a novel mathematical tool for the comparative analysis of classifiers when dealing with multi–class datasets.

Finally, the use of prediction probabilities instead of the confusion matrix opens new research directions to deepen prediction performance measures, like the MCP curve. Future work will focus on investigating the impact of prediction probabilities on performance metrics for classifiers.

REFERENCES

[1] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.
[2] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Comput. Biol. Chem.*, vol. 28, nos. 5–6, pp. 367–374, Dec. 2004.
[3] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
[4] C. J. Van Rijsbergen, *Information Retrieval*. Town, Malaysia: Butterworths, 1975.
[5] W. Peterson, T. Birdsall, and W. Fox, "The theory of signal detectability," *Trans. IRE Prof. Group Inf. Theory*, vol. 4, no. 4, pp. 171–212, Sep. 1954.
[6] J. A. Swets, "The relative operating characteristic in psychology," *Science*, vol. 182, no. 4116, pp. 990–1000, Dec. 1973.
[7] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.
[8] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[9] D. Ponce, L. G. M. de Andrade, C. D. Granado, A. Ferreiro-Fuentes, and R. Lombardi, "Development of a prediction score for in-hospital mortality in COVID-19 patients with acute kidney injury: A machine learning approach," *Sci. Rep.*, vol. 11, no. 1, p. 24439, Dec. 2021.
[10] L. Lu, B. Anderson, R. Ha, A. D'Agostino, S. L. Rudman, D. Ouyang, and D. E. Ho, "A language-matching model to improve equity and efficiency of COVID-19 contact tracing," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 43, Oct. 2021, Art. no. e2109443118.
[11] T. M. Bury, R. I. Sujith, I. Pavithran, M. Scheffer, T. M. Lenton, M. Anand, and C. T. Bauch, "Deep learning for early warning signals of tipping points," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 39, Sep. 2021, Art. no. e2106140118.
[12] M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 1, Jan. 2022, Art. no. e2110013119.
[13] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *Lancet Digit. Health*, vol. 1, no. 6, pp. e271–e297, Oct. 2019.
[14] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit. Lett.*, vol. 30, no. 1, pp. 27–38, 2009.
[15] J. Hernandez-Orallo, P. A. Flach, and C. Ferri, "A unified view of performance metrics: Translating threshold choice into expected classification loss," *J. Mach. Learn. Res.*, vol. 13, pp. 2813–2869, Oct. 2012.
[16] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, 2001.
[17] D. Mossman, "Three-way ROCs," *Med. Decis. Making*, vol. 19, no. 1, pp. 78–89, Jan. 1999.
[18] C. Ferri, J. Hernández-Orallo, and M. A. Salido, "Volume under the ROC surface for multi-class problems," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 2003, pp. 108–120.
[19] X. He and E. C. Frey, "The meaning and use of the volume under a three-class ROC surface (VUS)," *IEEE Trans. Med. Imag.*, vol. 27, no. 5, pp. 577–588, May 2008.
[20] E. Hellinger, "Die orthogonalvarianten quadratischer Formen von unendlich vielen variablelen," Ph.D. thesis, Dept. Math., Univ. Göttingen, Göttingen, Germany, 1907.
[21] E. Hellinger, "Neue begründung der theorie quadratischer Formen von unendlichvielen Veränderlichen," *J. Für Die Reine Und Angew. Math.*, vol. 1909, no. 136, pp. 210–271, Jul. 1909.
[22] G. Zhao, T. Vatanen, L. Droit, A. Park, A. D. Kostic, T. W. Poon, H. Vlamakis, H. Siljander, T. Härkönen, A. M. Hämäläinen, and A. Peet, "Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 30, pp. 6166–6175, Jul. 2017.
[23] M. Cartolano, N. Abedpour, V. Achter, T.-P. Yang, S. Ackermann, M. Fischer, and M. Peifer, "CaMuS: Simultaneous fitting and de novo imputation of cancer mutational signature," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, Dec. 2020.
[24] W. Peng, H. Deng, and A. Chen, "Using Hellinger and Bures metrics to construct two-dimensional quantum metric space for weather data fusion," *Inf. Fusion*, vol. 55, pp. 199–206, Mar. 2020.
[25] L. Wasserman, A. Ramdas, and S. Balakrishnan, "Universal inference," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 29, pp. 16880–16890, 2020.
[26] C. Zhu and F. Xiao, "A belief Hellinger distance for D–S evidence theory and its application in pattern recognition," *Eng. Appl. Artif. Intell.*, vol. 106, Nov. 2021, Art. no. 104452.
[27] N. D. Rochman, Y. I. Wolf, G. Faure, P. Mutz, F. Zhang, and E. V. Koonin, "Ongoing global and regional adaptive evolution of SARS-CoV-2," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 29, Jul. 2021, Art. no. e2104241118.
[28] I. Batzianoulis, F. Iwane, S. Wei, C. G. P. R. Correia, R. Chavarriaga, J. D. R. Millán, and A. Billard, "Customizing skills for assistive robotic manipulators, an inverse reinforcement learning approach with error-related potentials," *Commun. Biol.*, vol. 4, no. 1, p. 1406, Dec. 2021.
[29] C. Rosswog, C. Bartenhagen, A. Welte, Y. Kahlert, N. Hemstedt, W. Lorenz, M. Cartolano, S. Ackermann, S. Perner, W. Vogel, and J. Altmüller, "Chromothripsis followed by circular recombination drives oncogene amplification in human cancer," *Nature Genet.*, vol. 53, no. 12, pp. 1673–1685, Dec. 2021.
[30] Z. Duren, W. S. Lu, J. G. Arthur, P. Shah, J. Xin, F. Meschi, M. L. Li, C. M. Nemec, Y. Yin, and W. H. Wong, "Sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data," *Nature Commun.*, vol. 12, no. 1, p. 4763, Dec. 2021.
[31] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, no. 4, pp. 493–507, Dec. 1952.

[32] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[33] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, no. 1, pp. 99–109, 1943.
[34] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban, "The five factor model of personality and evaluation of drug consumption risk," in *Data Science*. Cham, Switzerland: Springer, 2017, pp. 231–242.
[35] D. Chicco, M. J. Warrens, and G. Jurman, "The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment," *IEEE Access*, vol. 9, pp. 78368–78381, 2021.

JESÚS S. AGUILAR-RUIZ worked as the Dean of the School of Engineering, Pablo de Olavide University, Seville, Spain, from 2005 to 2015, where he is currently a Full Professor in data analytics science and engineering. He has published over 250 articles in international conferences and journals. His research interests include data analytics and machine learning. He has been the Founder Editor-in-Chief of *BioData Mining* journal, from 2008 to 2014.

MARCIN MICHALAK was born in Mikolów, Poland, in 1981. He received the M.S. (Eng.) and Ph.D. degrees in computer science from the Silesian University of Technology (SUT), Gliwice, Poland, in 2005 and 2009, respectively.

Since 2012, he has been an Assistant Professor at SUT and a Senior Specialist with the Research Network Łukasiewicz—Institute of Innovative Technologies EMAG, Katowice, Poland. He has published over 90 scientific papers. His research interests include rough set theory, biclustering, machine learning, and data analysis.